SCIENCE AND TECHNOLOGY TEXT MINING: NEAR-EARTH SPACE

Dr. Ronald N. Kostoff, Office of Naval Research, 800 N. Quincy St., Arlington, VA 22217; Phone: 703-696-4198; Fax: 703-696-4274; Internet: kostofr@onr.navy.mil

Mr. Henry J. Eberhart, Naval Air Warfare Center (Ret'd), China Lake, CA

Mr. Darrell Ray Toothman, RSIS, Inc., McLean, VA

ABSTRACT

Database Tomography is a patented system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and performing phrase proximity analyses. This report shows how Database Tomography can be used to derive technical intelligence from the published literature.

One potential application of Database Tomography is to obtain the thrusts and interrelationships of a technical field from papers published in the literature within that field. This paper provides an application of Database Tomography to analysis of Near-Earth Space (NES) science and technology, to support the Navy's activities in this region of space.

A database of relevant NES articles was analyzed to produce characteristics and key features of the NES field. The recent prolific NES authors, the journals prolific in NES papers, the prolific institutions in NES, the prolific keywords specified by the authors, and the authors whose works are cited most prolifically, as well as the particular papers/ journals/ institutions cited most prolifically, are identified. The pervasive themes of NES are identified through multi-word phrase analyses of the database. A phrase proximity analysis of the database shows the relationships among the pervasive themes, and the relationships between the pervasive themes and sub-themes.

(*The views in this paper are solely those of the authors and do not represent the views of the Department of the Navy, any of its components, or of RSIS, Inc*).

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)
21-07-2003

2. REPORT TYPE
Technical

3. DATES COVERED (FROM - TO)
xx-xx-1996 to xx-xx-1998

4. TITLE AND SUBTITLE
SCIENCE AND TECHNOLOGY TEXT MINING
NEAR-EARTH SPACE
Unclassified

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kostoff, Ronald N ;
Eberhart, Henry J ;
Toothman, Darrell R ;

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APUBLIC RELEASE
,

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Database Tomography is a patented system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and performing phrase proximity analyses. This report shows how Database Tomography can be used to derive technical intelligence from the published literature. One potential application of Database Tomography is to obtain the thrusts and interrelationships of a technical field from papers published in the literature within that field. This paper provides an application of Database Tomography to analysis of Near-Earth Space (NES) science and technology, to support the Navy's activities in this region of space. A database of relevant NES articles was analyzed to produce characteristics and key features of the NES field. The recent prolific NES authors, the journals prolific in NES papers, the prolific institutions in NES, the prolific keywords specified by the authors, and the authors whose works are cited most prolifically, as well as the particular papers/ journals/ institutions cited most prolifically, are identified. The pervasive themes of NES are identified through multi-word phrase analyses of the database. A phrase proximity analysis of the database shows the relationships among the pervasive themes, and the relationships between the pervasive themes and sub-themes.

15. SUBJECT TERMS
space vehicle; low earth orbit; satellite; text mining, information retrieval, communications; navigation; bibliometrics; computational linguistics

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE
Unclassified	Unclassified	Unclassified

17. LIMITATION OF ABSTRACT
Same as Report (SAR)

18. NUMBER OF PAGES
36

19. NAME OF RESPONSIBLE PERSON
Kostoff, Ronald
kostofr@onr.navy.mil

19b. TELEPHONE NUMBER
International Area Code
Area Code Telephone Number
703696-4198
DSN
-

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39.18

KEYWORDS: space vehicle; low earth orbit; satellite; text mining, information retrieval, communications; navigation; bibliometrics; computational linguistics.

BACKGROUND

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a concomitant increase in the need for scientific and technical intelligence to insure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While there is no substitute for direct human intelligence gathering, there have become available many techniques that can support and complement direct human intelligence gathering. In particular, techniques that identify, select, gather, cull, and interpret large amounts of technological information semi-autonomously can expand greatly the capabilities of human beings for performing technical intelligence.

This report shows how Database Tomography (1-5) can be used to derive technical intelligence from the published literature. Database Tomography is a patented system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and performing phrase proximity analyses. Phrase frequency analysis provides the pervasive themes of a database, and the phrase proximity analysis provides the relationships among the pervasive themes, and between the pervasive themes and sub-themes.

One potential application of Database Tomography is to obtain the thrusts and interrelationships of a technical field from papers published in the literature within that field. This paper provides an application of Database Tomography to analysis of Near-Earth Space (NES) science and technology, to support the Navy's activities in this region of space.

In particular, NES, as defined by the authors for this study, consists of unmanned space vehicle operation in low earth orbit focusing on earth-related activities such as communications, navigation, and observations. For the most part, planetary observations and other non-earth oriented activities were included only if they illuminated science and technology issues that could be of use to the earth-oriented activities.

To execute the study in this report, a database of relevant NES articles is generated using a unique and powerful search approach (6, 7), and the database is analyzed to produce characteristics and key features of the NES field. The recent prolific NES authors, the journals prolific in NES papers, the prolific institutions in NES, the prolific keywords specified by the authors, and the authors whose works are cited most prolifically, as well as the particular papers/ journals/institutions cited most prolifically, are identified. The pervasive themes of NES are identified through multi-word phrase analyses of the database. A phrase proximity analysis of the database shows the relationships among the pervasive themes, and the relationships between the pervasive themes and subthemes.

What is the importance of applying Database Tomography to a topical field such as NES? Database Tomography provides a map of the field of interest and, analogous to ordinary roadmaps, serves as a structured guide to reach a specific destination efficiently. Suppose one wants to understand the limitations of existing NES science and technology, and perhaps identify promising avenues for improvement. One could start with hit-or-miss literature searches or randomized personal contacts, or one could start with Database Tomography.

Database Tomography would identify the main intellectual thrust areas in NES, and the relationships among those thrust areas. As part of the analysis output, the main NES techniques conceptualized and employed would be identified. The major journals associated with each thrust area and technique would be identified, the major authors for each technique and thrust area would be identified, and the major institutions and countries associated with each technique and thrust area would be identified. The ancillary techniques and the science and technology areas which could support and improve a technique or thrust area would be identified, and conversely techniques or thrust areas that could be impacted by a given technique would be identified.

The map, then, provides a comprehensive overview of the full picture, and allows specific starting points to be chosen rationally for more detailed investigations into a topic of interest. It does not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area in order to make a substantial contribution to the understanding or the advancement of this topical area, but allows these detailed efforts to be executed more efficiently.

DATABASE GENERATION

The key step in the NES literature analysis is the generation of the database. For the present study, the database consists of selected journal and conference proceeding abstracts (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the Science Citation Index (SCI) and the Engineering Compendex (EC). The SCI accesses about 3200 journals (mainly in physical and life sciences basic research) and the EC accesses about 2600 journals and conference proceedings (mainly in applied research and technology). In the SCI and EC, the title, keyword, and abstract fields were searched using keywords relevant to NES, although different procedures were used to search the title and abstract fields (6). The resultant abstracts were culled to those relevant to NES.

The search was performed with the aid of two powerful Database Tomography tools (multi-word phrase frequency analysis and phrase proximity analysis) using the process of Simulated Nucleation (6, 7). An initial database of titles, keywords, and abstracts was created from a core of papers known to be highly relevant to NES. Phrase frequency and proximity analyses were performed on this textual database. The high frequency single, double, and triple word phrases obviously relevant to NES, and their Boolean combinations, were then used as search terms in the SCI and EC databases. The process was repeated on the new database of titles, keywords, and abstracts which were found. A few more iterations were performed until convergence was obtained. One value of utilizing Simulated Nucleation is that the search terms are obtained from the words of the authors in the SCI and EC databases, not by guessing on the part of the searcher.

PROLIFIC AUTHORS

The author field was separated from the database, and a frequency count of author appearances was made. The most prolific authors follow, in order of decreasing publications. A caveat is in order here. The journals and conference proceedings searched were limited to those in the SCI and EC. Relevant articles in other journals or conference proceedings were not included. Books or major reports were not included. The keywords used were a finite set of the authors' discretion, and undoubtedly overlooked some relevant articles in NES. The time frame of the articles was 1993-mid 1996. Thus, there may be excellent researchers writing in the field of NES who were omitted from the following list due to the finite selection process, and the authors' apologies are extended to anyone who falls into this category.

In the SCI database results, there were 12453 different authors. There were 18474 author listings (the occurrence of each author's name on a paper is defined as an author listing) and 5481 papers retrieved, yielding an average of 3.37 authors per paper. While this may seem high at first glance, a detailed examination of the results shows that these space-related papers tend to have large collaborative groups. Especially in the space experiments, large facilities, efforts, and costs are involved, and typically many different experiments are performed. Many of the efforts tend to involve multiple disciplines. The large number of collaborators reflects this diversity.

In the EC database results, there were 14036 different authors. There were 17754 author listings and 5617 papers retrieved, yielding an average of 3.16 authors per paper.

TABLE 1 - MOST PROLIFIC AUTHORS

<u>SCI</u>

WATERS-JW 32;
FROIDEVAUX-L 21;
HAYS-PB 18;
SHEPHERD-GG 18;
NEWELL-PT 17;
SKINNER-WR 16;
CRACKNELL-AP 15;
MILLER-AJ 15;
BURRAGE-MD 15;
RUSSELL-JM 14;
GILLE-JC 13;
VAROTSOS-CA 13;
BURKE-WJ 12;

HARWOOD-RS 12;
FISHBEIN-EF 12;
WINNINGHAM-JD 12

<u>EC</u>

EVANS B.G.,16;
KATO SHUZO,16;
KONDRAT'EV K.YA.,14;
COAKLEY F.P.,12;
BROWN ALISON,12;
KUBOTA SHUJI,12;
BERTHOD MARC,12;
HADJITHEODOSIOU M.H.,11;
MODI V.J.,9;
KWATRA S.C.,8;
DE GAUDENZI RICCARDO,8;
ELHAKEEM A.K.,8;
ALFANO SALVATORE,8;

CODE: THE NUMBER FOLLOWING EACH AUTHOR'S NAME REPRESENTS THE NUMBER OF PAPERS AUTHORED OR CO-AUTHORED IN THE LITERATURE DATABASE.

There is no overlap between the two lists, showing the distinct separation between those performers doing space research and space technology.

PROLIFIC JOURNALS

A similar process was used to develop a frequency count of journal appearances. Similar limitations to those mentioned above apply to the journals, and similar apologies are extended to journals not listed. In the SCI database, there were 628 different journals represented. The most prolific journals follow in order of decreasing frequency. While many disciplines are represented in this table, there seems to be large representation from the environmental sensing field in the SCI journals. The top 2% of SCI journals, in terms of NES articles published, accounted for 41% of all NES articles. It should be noted here that the databases accessed are unclassified, and the conclusions reached apply to the unclassified literature only.

In the EC database, there were over 1000 journals and conference proceedings listed. As contrasted to the SCI results, the highest frequency EC publication sources were conference proceedings.

TABLE 2 - MOST PROLIFIC JOURNALS

SCI

ADVANCES IN SPACE RESEARCH,335;
INTERNATIONAL JOURNAL OF REMOTE SENSING,314;
JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES,254;
JOURNAL OF GEOPHYSICAL RESEARCH-OCEANS,228;
GEOPHYSICAL RESEARCH LETTERS,209;
JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS,205;
JOURNAL OF CLIMATE,133;
JOURNAL OF APPLIED METEOROLOGY,132;
REMOTE SENSING OF ENVIRONMENT,108;
MONTHLY WEATHER REVIEW,92;

EC

PROCEEDINGS OF SPIE -THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING,492;
INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS),414;
IEE CONFERENCE PUBLICATION,251;
ADVANCES IN THE ASTRONAUTICAL SCIENCES,181;
INTERNATIONAL JOURNAL OF REMOTE SENSING,142;
ISSLEDOVANIE ZEMLI IZ KOSMOSA,125;
IEE COLLOQUIUM (DIGEST),106;
BETTER UNDERSTANDING OF EARTH ENVIRONMENT INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS),93;
PROCEEDINGS OF ION GPS,93;
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING,84;
CONFERENCE PROCEEDINGS -EUROPEAN MICROWAVE

CONFERENCE,73;
REMOTE SENSING OF ENVIRONMENT,72;
INTERNATIONAL JOURNAL OF SATELLITE COMMUNICATIONS,67;
ELEKTROSVYAZ,67;

CODE: THE NUMBER FOLLOWING EACH JOURNAL REPRESENTS THE NUMBER OF PAPERS IN THE LITERATURE DATABASE PUBLISHED IN THE JOURNAL

Two journals are shared by the lists (INTERNATIONAL JOURNAL OF REMOTE SENSING, REMOTE SENSING OF ENVIRONMENT), both focused on remote sensing. Most of the archives listed from the SCI are fundamental science journals, while most of the archives listed from the EC are conference proceedings.

PROLIFIC INSTITUTIONS

A similar process was used to develop a frequency count of institutional address appearances, and similar apologies are extended to institutions not listed. There were 10435 different organizations listed in the SCI author address organizations. The most prolific institutions follow in order of decreasing frequency in Table 3. It should be noted that many different organizational components may be included under the single organizational heading (e.g., Harvard Univ could include the Chemistry Department, Biology Department, Physics Department, etc.). Lack of space precluded printing out the components under the organizational heading. The top 2% of all SCI organizations accounted for 36% of total addresses listed.

In the EC database, there were 2747 different organizations listed from 73 countries. However, the EC database is less uniform than the SCI database, and what appears to be separate organizations is in many cases different terminology for the same organization. For example, NASA Goddard Space Flight Center had 13 separate terminology listings with a total count of 68, and the Naval Research Laboratory had 10 separate listings with a total count of 39. Thus, the number of distinct organizations is probably closer to 2000.

For the SCI, the total NASA frequency count, including a few separate NASA center listings but not the Jet Propulsion Lab (JPL) run by the California Institute of Technology, was about 670. The total U. S. Navy count, including a few separate component listings, was about 100, or about 15% of the NASA count. For the EC, the total NASA frequency count, including all the separate center listings but not JPL, was about 133, and the total U. S. Navy count, including the separate component listings, was about 66, or about 50% of the NASA count.

TABLE 3 - MOST PROLIFIC INSTITUTIONS

SCI

NASA,667;
NOAA,313;
UNIV-COLORADO,173;
CALTECH,148;
RUSSIAN-ACAD-SCI,120;
UNIV-MARYLAND,106;
USN,96;
UNIV-WISCONSIN,93;
NATL-CTR-ATMOSPHER-RES,91;
UNIV-CALIF-SAN-DIEGO,89;
JOHNS-HOPKINS-UNIV,88;
UNIV-TEXAS,86;
UNIV-WASHINGTON,82;
JET-PROP-LAB,72;

EC

CALIFORNIA INST OF TECHNOLOGY,99;
NASA/GODDARD SPACE FLIGHT CENT,68;
UNIV OF CALIFORNIA,47;
MITRE CORP,42;
UNIV OF SURREY,40;
NAVAL RESEARCH LAB,39;
COMSAT LAB,39;
AEROSPACE CORP,37;
UNIV OF COLORADO,31;
NTT RADIO COMMUNICATION SYSTEMS LAB,29;

EUROPEAN SPACE AGENCY,29;
NATL AERONAUTICS AND SPACE ADMINISTRATION,26;
UNIV OF MARYLAND,25;
RUSSIAN ACAD OF SCIENCES,24;
UNIV OF CALGARY,23;
STANFORD UNIV,23;
JET PROPULSION LAB,22;
PHILLIPS LAB,22;
ISRO SATELLITE CENT,22;
JOHNS HOPKINS UNIV,20;

CODE: THE NUMBER FOLLOWING EACH INSTITUTION REPRESENTS THE NUMBER OF TIMES A NAME OF A REPRESENTATIVE FROM THAT INSTITUTION APPEARS AS AN AUTHOR OR CO-AUTHOR IN THE LITERATURE DATABASE

No industrial organizations are listed in the SCI list shown, whereas four industrial organizations are listed in the EC list shown.

PROLIFIC COUNTRIES

The most prolific countries follow in order in Table 4. There were 105 different countries listed in the SCI results. The top 2% of SCI countries accounted for 63% of total addresses. The dominance of a handful of countries is clearly evident. In the EC database, there were 73 countries listed, and the same handful of countries was dominant. However, in the SCI database, the ratio of USA listings to the sum of the ten next highest country listings was 1.35, while in the EC database this
ratio was 0.9. Thus, while the USA dominates both science (as represented by SCI counts) and technology (as represented by EC counts), the dominance appears noticeably greater in science than technology. Again, all these conclusions apply only to the open literature data, not classified or company proprietary data.

TABLE 4 - MOST PROLIFIC COUNTRIES

SCI

USA,5266;

UNITED KINGDOM,660;
FRANCE,614;
JAPAN,549;
CANADA,476;
GERMANY,417;
RUSSIA,370;
ITALY,274;
AUSTRALIA,207;
INDIA,203;
NETHERLANDS,127;
SWEDEN,119;
NORWAY,92;
SPAIN,85;
PEOPLES-R-CHINA,71;
DENMARK,62;
FINLAND,54;
SWITZERLAND,52;

EC

USA,2354;
JAPAN,479;
UK,445;
FRANCE,342;
GERMANY,275;
CANADA,272;
ITALY,230;
RUSSIA,209;
INDIA,143;
NETHERLANDS,125;
AUSTRALIA,103;
CHINA,77;
SPAIN,54;
SWEDEN,42;
NORWAY,39;
SWITZERLAND,33;
TAIWAN,32;
ISRAEL,29;

DENMARK,28;
BELGIUM,25;
FINLAND,24;
BRAZIL,24;
GREECE,21;
UKRAINE,16.

PROLIFIC CITATIONS

The citations in all 5481 SCI papers were aggregated into a file of 140,662 entries. The authors, specific papers, years, journals, and countries most frequently cited were identified, and are presented in order of decreasing frequency in the following tables. Each table format has the entry followed by the number of times the entry was cited.

There were 42094 different authors cited. 58% of the cited authors were cited once, and 16% of the cited authors were cited twice. A relatively few percent received large numbers of citations. However, the most cited authors, while prolific, are not the most prolific authors, and vice versa. For example, the three most highly cited authors (Rossow, Sellers, Tucker) ranked numbers 35, 50, and 100, respectively, in prolificity. The three most prolific authors (Waters, Froidevaux, Hays) ranked numbers 70, 1400, and 95, respectively, in citability. Part of this difference may be due to the time lag between the highly cited authors' productivity at the time their highly cited papers were written and their productivity today, as well as the phase in their career of the prolific authors. More probably, the time required to produce a seminal highly cited paper does not allow overly high volumes of papers to be produced.

There were 93194 different papers cited. 81% of cited papers received only one citation and 10% received two citations. Relatively few papers were highly cited. From the citation year results, the most recent papers are the most highly cited. This reflects a dynamic rapidly evolving field of research. There were 28740 different journals and other sources cited. 75% of sources were cited only once and 11% were cited twice. Relatively few sources were highly cited. There is a stronger correlation between highly prolific and cited journals than between highly prolific and cited authors. The time span over which a journal develops and maintains a reputation for high quality is long compared to the gap between publication and citation, and one should expect that in the steady state the prolific journals would also be the most highly cited. JGR and its serial journals (ATMOSPHERES, OCEANS, etc.) and International Journal of Remote Sensing tended to be highly prolific and cited, but the Journal of Atmospheric Science was highly cited but not prolific. One possible explanation is that the published papers are slightly more applied than some of their references, and if the Journal of Atmospheric Science tends to contain mainly very fundamental papers, it would serve mainly as a citing source, but not a publishing source.

The reader interested in researching the NES field would be well-advised to, first, obtain the highly-cited papers listed here and, second, peruse those sources that are highly cited and/or that contain large numbers of recently published NES papers.

TABLE 5 - MOST CITED AUTHORS - SCI ONLY

ROSSOW-WB,366;
SELLERS-PJ,357;
TUCKER-CJ,342;
STEPHENS-GL,214;
RAMANATHAN-V,198;
GOWARD-SN,192;
NEWELL-PT,184;
PRICE-JC,177;
SPENCER-RW,169;
SMITH-WL,160;
GORDON-HR,158;
LIU-WT,157;
CARTWRIGHT-DE,155;
CESS-RD,155;
KAUFMAN-YJ,153;
LOCKWOOD-M,152;
HOLBEN-BN,151;
JUSTICE-CO,151;
CHENEY-RE,149;
TOWNSHEND-JRG,147;
ULABY-FT,144;
ARKIN-PA,134;
FU-LL,130;

WARREN-SG,130;

TABLE 6 - MOST CITED PAPERS - SCI ONLY

B-AM-METEOROL-SOC V72,1991,ROSSOW-WB,111
J-GEOPHYS-RES V90,1985,MCCLAIN-EP,76
J-GEOPHYS-RES-ATMOS V97,1992,SELLERS-PJ,72
INT-J-REMOTE-SENS V6,1985,SELLERS-PJ,69
INT-J-REMOTE-SENS V7,1986,HOLBEN-BN,69
SCIENCE V227,1985,TUCKER-CJ,66
INT-J-REMOTE-SENS V12,1991,PRINCE-SD,63
J-CLIMATOL V1,1988,REYNOLDS-RW,63
SCIENCE V243,1989,RAMANATHAN-V,62
INT-J-REMOTE-SENS V6,1985,JUSTICE-CO,58
B-AM-METEOROL-SOC V64,1983,SCHIFFER-RA,57
J-GEOPHYS-RES V96,1991,NEWELL-PT,56

TABLE 7 - MOST CITED YEARS - SCI ONLY

1991,12053;
1992,11766;
1990,11394;
1993,10212;
1989,9611;
1988,8104;
1987,7433;
1994,7166;
1986,6777;
1985,5844;
1984,5258;
1983,4521;
1982,4022;
1981,3850;
1980,3185;
1979,2780;
1995,2647;
1978,2519;
1977,2149;

1976,2008;

TABLE 8 - MOST CITED JOURNALS - SCI ONLY

J-GEOPHYS-RES,10109;
J-ATMOS-SCI,3796;
GEOPHYS-RES-LETT,3213;
INT-J-REMOTE-SENS,2955;
J-GEOPHYS-RES-OCEANS,2928;
ASTROPHYS-J,2800;
J-GEOPHYS-RES-ATMOS,2473;
MON-WEATHER-REV,2335;
NATURE,2080;
REMOTE-SENS-ENVIRON,2012;
J-APPL-METEOROL,1858;
SCIENCE,1586;
J-CLIMATE,1544;
B-AM-METEOROL-SOC,1415;
J-PHYS-OCEANOGR,1379;
IEEE-T-GEOSCI-REMOTE,1228;
APPL-OPTICS,1184;
PLANET-SPACE-SCI,1119;
Q-J-ROY-METEOR-SOC,950;

PROLIFIC KEYWORDS

Table 9 contains the frequency distributions of the keywords associated with each paper. The EC numbers are about an order of magnitude higher than the SCI numbers, but the inclusion of classification categories in the EC keyword listings probably accounts for a substantial part of this difference.

The SCI keyword focus appears to be on environmental science characteristics (OCEAN, CLIMATE, TEMPERATURE, CIRCULATION, PRECIPITATION, CLOUD), whereas the EC keywords evidence a strong focus on satellite communications (SATELLITE COMMUNICATION SYSTEMS. COMMUNICATION SATELLITES, COMMUNICATION CHANNELS). Again, this EC conclusion applies only to the unclassified and non-proprietary literature.

TABLE 9 - MOST PROLIFIC KEYWORDS

<u>SCI</u>

MODEL,250;
SATELLITE DATA,146;
SATELLITE,139;
OCEAN,106;
CLIMATE,98;
VARIABILITY,84;
GALAXIES,83;
SURFACE,81;
DYNAMICS,80;
TEMPERATURE,75;
STARS,74;
CIRCULATION,72;
PRECIPITATION,71;
RADIATION,67;
CLOUD,61;
REMOTE SENSING,59;
VEGETATION,58;
SYSTEM,57;
ATMOSPHERE,56;
EVOLUTION,55;
SUN,51;
GENERAL-CIRCULATION MODEL,50;
SCATTERING,47;
GEOSAT
ALTIMETRY,46;
PLASMA,46;
CONVECTION,45;
FIELD,45;
SEA,44;
WATER,44;
SEA-SURFACE TEMPERATURE,43;
PARAMETERIZATION,42;
MAGNETOSPHERE,42;

REGION,42;
WAVES,42;
AVHRR DATA,40;
CLOUDS,40;
SIMULATION,40;
OZONE,39;
REFLECTANCE,39;
RAINFALL,39;
CLASSIFICATION,38;
IMAGERY,38;
TRANSPORT,38;
RETRIEVAL,38;
CALIBRATION,37;
AVHRR,37;
ULTRAVIOLET,37;
BOUNDARY-LAYER,36;
GLOBAL POSITIONING SYSTEM,35;
ABSORPTION,34;
IONOSPHERE,34;
EMISSION,34;
PACIFIC,34;
PHOTOSYNTHESIS,34;
MODELS,34;
SPECTRA,34;
SYSTEMS,34;
WATER-VAPOR,34;
EARTH,33;
COVER,32;
PROFILES,30;

EC

SATELLITES,1770;
SATELLITE COMMUNICATION SYSTEMS,1105;
MATHEMATICAL MODELS,910;
REMOTE SENSING,708;
COMMUNICATION
SATELLITES,599;

ALGORITHMS,557;
COMMUNICATION CHANNELS (INFORMATION THEORY),433;
ORBITS,422;
SATELLITE LINKS,395;
COMPUTER SIMULATION,392;
PERFORMANCE,341;
GLOBAL POSITIONING
SYSTEM,311;
DATA ACQUISITION,307;
TELECOMMUNICATION SERVICES,285;
RADIOMETERS,257;
GEOSTATIONARY SATELLITES,251;
TRACKING (POSITION),243;
SENSORS,242;
DATA COMMUNICATION SYSTEMS,236;
ANTENNAS,236;
MOBILE TELECOMMUNICATION SYSTEMS,235;
IMAGE PROCESSING,227;
COMPUTER SOFTWARE,223;
TELECOMMUNICATION NETWORKS,219;
SATELLITE GROUND STATIONS,216;
SYNTHETIC APERTURE RADAR,213;
SPACECRAFT,209;

PERVASIVE THEMES

To obtain pervasive themes, single, double, and triple word phrases from the text of the database were identified, and the high frequency high technical content words were identified as the pervasive themes. In this particular exercise, the database was split into two parts, and the analysis was done on each part. The titles of the papers were put into a separate database, and the multiword frequency analysis was performed. The abstracts of the papers constituted a separate database as well. Since the highest frequency phrases from the title and abstract databases were very similar, only raw data outputs from the abstract database will be presented here.

In the following tables, the number preceding the phrase is the frequency of appearance of the phrase in the database. From a global perspective, the SCI database portrays the major applications of NES to be remote sensing of the sea and earth surface from satellites using SAR and high resolution radiometry to obtain temperatures and ice information, radiation budgets, and vegetation and crop information, as well as navigation using GPS. The EC database confirms the SCI thrusts listed, but in addition shows a major technological emphasis on satellite communications and associated hardware.

TABLE 10 - ABSTRACT DOUBLE WORD FREQUENCIES

SCI

579 SATELLITE DATA;
541 REMOTE SENSING;
413 SEA SURFACE;
400 WATER VAPOR;
396 SURFACE TEMPERATURE;
312 HIGH RESOLUTION;
285 MAGNETIC FIELD;
250 SATELLITE IMAGERY;
236 SATELLITE OBSERVATIONS;
234 GLOBAL POSITIONING;
231 POSITIONING SYSTEM;
216 DATA SETS;
210 TIME SERIES;
207 BOUNDARY LAYER;
207 SEA LEVEL;
205 DATA SET;
183 LAND SURFACE;
182 RESOLUTION RADIOMETER;
181 RADIATIVE TRANSFER;
177SATELLITE IMAGES;
175 SPATIAL RESOLUTION;
174 SEA ICE;
166 SYSTEM GPS;
166 TOPEX POSEIDON;
165 SATELLITE MEASUREMENTS;
163 RADIATION BUDGET;
161 TOTAL OZONE;

146 INTERNATIONAL SATELLITE;
146 NORTHERN HEMISPHERE;
144 CLIMATOLOGY PROJECT;
142 VEGETATION INDEX;
137 CLOUD COVER;
135 BRIGHTNESS TEMPERATURES;
132 UPPER ATMOSPHERE;
128 WIND SPEED;
126 SOIL MOISTURE;
126 SOLAR RADIATION;

EC

587 REMOTE SENSING;
377 SATELLITE COMMUNICATIONS;
257 SATELLITE COMMUNICATION;
238 MOBILE SATELLITE;
238 SATELLITE DATA;
220 GLOBAL POSITIONING;
211SATELLITE SYSTEM;
210 HIGH RESOLUTION;
204 SATELLITE SYSTEMS;
202 POSITIONING SYSTEM;
186 DOUBLE PRIME;
172 COMMUNICATION SYSTEMS;
170 MULTIPLE ACCESS;
165 EARTH ORBIT;
163 LOW EARTH;
153 SEA SURFACE;
145 SYNTHETIC APERTURE;
134 SYSTEM GPS;
133 SURFACE TEMPERATURE;
127 SATELLITE IMAGES;
125 APERTURE RADAR;
120 SIGNAL PROCESSING;
120 SPATIAL RESOLUTION;
119 SPACE AGENCY;
118 SATELLITE IMAGERY;

112 BIT ERROR;
112 SEA ICE;
110 UNITED STATES;
108 LOW COST;
105 COMMUNICATIONS SYSTEMS;
105 NAVIGATION SYSTEM;
103 COMMUNICATION SYSTEM;
101 REMOTELY SENSED;
100 CONTROL SYSTEM;

CODE: THE NUMBER FOLLOWING EACH WORD PAIR REPRESENTS THE NUMBER OF TIMES THE WORD PAIR APPEARED IN ALL THE ABSTRACTS OF THE LITERATURE DATABASE

TABLE 11 - ABSTRACT TRIPLE WORD FREQUENCIES

SCI

227 GLOBAL POSITIONING SYSTEM;
193 SEA SURFACE TEMPERATURE;
178 HIGH RESOLUTION RADIOMETER;
165 POSITIONING SYSTEM GPS;
118 RESOLUTION RADIOMETER AVHRR;
106 SPECIAL SENSOR MICROWAVE;
101 ATMOSPHERE RESEARCH SATELLITE;
97 INTERNATIONAL SATELLITE CLOUD;
97 SATELLITE CLOUD CLIMATOLOGY;
97 SENSOR MICROWAVE IMAGER;
97 SPATIAL AND TEMPORAL;
93 EARTH RADIATION BUDGET;
89 CLOUD CLIMATOLOGY PROJECT;
85 DIFFERENCE VEGETATION INDEX;
84 NORMALIZED DIFFERENCE VEGETATION;
80 SURFACE TEMPERATURE SST;
75 SYNTHETIC APERTURE RADAR;
70 GENERAL CIRCULATION MODEL;

70 SATELLITE REMOTE SENSING;
64 MICROWAVE IMAGER SSM;
63 RADIATIVE TRANSFER MODEL;
63 VEGETATION INDEX NDVI;
60 OCEANIC AND ATMOSPHERIC;
59 RADIATION BUDGET EXPERIMENT;
52 SPACE AND TIME;
51 LANDSAT THEMATIC MAPPER;

<u>EC</u>

193 GLOBAL POSITIONING SYSTEM;
125 LOW EARTH ORBIT;
122 SYNTHETIC APERTURE RADAR;
87 EUROPEAN SPACE AGENCY;
75 SATELLITE COMMUNICATION SYSTEMS;
68 HIGH RESOLUTION RADIOMETER;
67 REMOTE SENSING SATELLITE;
60 DIVISION MULTIPLE ACCESS;
60 REMOTE SENSING DATA;
59 BIT ERROR RATE;
58 SEA SURFACE TEMPERATURE;
54 MOBILE SATELLITE COMMUNICATIONS;
54 RESOLUTION RADIOMETER AVHRR;
53 FIELD OF VIEW;
53 LANDSAT THEMATIC MAPPER;
51 EARTH ORBIT LEO;
51 REMOTELY SENSED DATA;
51 SATELLITE REMOTE SENSING;
49 EQUATIONS OF MOTION;
49 LAND MOBILE SATELLITE;
49 SATELLITE COMMUNICATIONS SYSTEMS;
48 THEMATIC MAPPER TM;
44 EARTH OBSERVING SYSTEM;
42 SATELLITE COMMUNICATION SYSTEM;
39 NORMALIZED DIFFERENCE VEGETATION;
38 DIGITAL SIGNAL PROCESSING;
37 DIFFERENCE VEGETATION INDEX;

37 MOBILE SATELLITE COMMUNICATION;
36 DATA RELAY SATELLITE;
35 PHASE SHIFT KEYING;
35 SPECIAL SENSOR MICROWAVE;
32 DIRECT BROADCAST SATELLITE;

TABLE 12 - ABSTRACT SINGLE WORD FREQUENCIES

SCI

6268 DATA;
5452 SATELLITE;
3000 SURFACE;
2995 MODEL;
2145 CLOUD;
2028 MEASUREMENTS;
1903 SYSTEM;
1842 OBSERVATIONS;
1742 TWO;
1683 TEMPERATURE;
1678 TIME;
1652 OBSERVED;
1506 WATER;
1489 FIELD;
1420 HIGH;
1377ANALYSIS;
1372 KM;
1350 GLOBAL;
1319 ATMOSPHERIC;
1309 REGION;
1258 SEA;
1253 SPACE;
1237 STUDY;
1139 DEGREES;
1104 SOLAR;
1098 METHOD;
1094 BASED;
1082 FOUND;

1081 LARGE;
1073 ICE;
1065 SATELLITES;
1062 PAPER;
1009 EARTH;
1008 RESOLUTION;
1000 MODELS;
962 RADIATION;
943 DERIVED;
938 OCEAN;
928 CURRENT;
925 SPATIAL;
899 ATMOSPHERE;
893 COMPARED;
890 MEAN;
883 VALUES;
880 WIND;
857 DISTRIBUTION;
856 IMAGES;
848 LOW;
846 RANGE;
836 PRESENTED;
833 CLOUDS;
807 CHANGES;
800 ENERGY;
800 ERROR;
792 SIMILAR;
789 VARIATIONS;
788 RADAR;
787 SYSTEMS;
785 TOTAL;
779 EFFECTS;
777 CONDITIONS;
774 REGIONS;
764 ESTIMATES;
762 INFORMATION;
757 PERIOD;
731 MEASURED;

EC

7119 SATELLITE;
4986 DATA;
4654 SYSTEM;
2268 SYSTEMS;
1788 SATELLITES;
1777 SPACE;
1626 PERFORMANCE;
1605 TWO;
1563 MODEL;
1448 METHOD;
1438 HIGH;
1427 TIME;
1375 BASED;
1373 DESIGN;
1358 GPS;
1356 COMMUNICATIONS;
1311 CONTROL;
1283 SURFACE;
1248 MEASUREMENTS;
1241 EARTH;
1233 ANALYSIS;
1122 NETWORK;
1114 INFORMATION;
1086 POWER;
1032 ORBIT;
986 ANTENNA;
979 COMMUNICATION;
973 IMAGES;
972 IMAGE;
967 LOW;
953 FREQUENCY;
941 APPLICATIONS;
936 DEVELOPMENT;
932 DIGITAL;
925 STUDY;

894 GROUND;
890 SIGNAL;
860 ACCURACY;
856 GLOBAL;
833 PHASE;
812 REMOTE;
799 TEMPERATURE;
792 PROCESSING;
791 CHANNEL;
785 FIELD;
783 RADAR;
768 ALGORITHM;
766 ERROR;
759 MOBILE;
754 RESOLUTION;
743 RANGE;
737 PARAMETERS;
729 TECHNIQUE;
714 OPTICAL;
714 SPACECRAFT;
711 DEGREE;
702 TRANSMISSION;
696 LARGE;
693 TEST;
686 NUMBER;
671 EFFECTS;
662 SPECTRAL;
661 COMPARED;
659 NAVIGATION;
652 RATE;
648 SMALL;
647 SENSING;
647 SOLAR;
644 ORDER;
638 METHODS;
623 NETWORKS;
611 ATMOSPHERIC;
601 MINUS;

597 STRUCTURE;
585 AREAS;

THEME RELATIONSHIPS

To obtain the theme and sub-theme relationships, a phrase proximity analysis is performed about each theme phrase. Typically, forty to sixty multi-word phrase themes are selected from a multi-word phrase analysis of the type shown above. For each theme phrase, the frequencies of phrases within +-50 words of the theme phrase for every occurrence in the full text are computed. A phrase frequency dictionary is constructed which shows the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each phrase frequency dictionary (hereafter called cluster) yield those sub-themes closely related to the main cluster theme.

Then, threshold values are assigned to the numerical indices. These indices are used to filter out the most closely related phrases to the cluster theme (e.g., see the example (TABLE 13- REMOTE SENSING-ABSTRACT DATABASE) following this section for part of a typical filtered cluster from the study). Because of space limitations in this document, only one theme, REMOTE SENSING, was chosen for the phrase proximity analysis. It was high frequency in both the abstracts and titles, and is a central theme for the utilization of near-earth space. The full text database was split into two databases. One was the abstract narrative, and it was hoped that performing the phrase proximity analysis on this database would yield mainly topical theme relationships. The other database consisted of records (one for each published paper) containing four fields: author(s), title, journal name, author(s) institutional address(es). It was hoped that performing the phrase proximity analysis on this database would yield not only topical theme relationships from the proximal title words, but also relationships between technical themes and authors, journals, and institutions.

TABLE 13

THEME WORD "REMOTE SENSING" - ABSTRACT DATABASE - SORT BY Ij

Cij	Ci	Ii (Cij/Ci)	Ij (Cij/Cj)	Eij (Ii*Ij)	CLUSTER MEMBER
752	10659	0.071	0.703	0.0496	DATA
575	12082	0.048	0.537	0.0256	SATELLITE REMOTE
260	4042	0.064	0.243	0.0156	SURFACE
207	1398	0.148	0.193	0.0286	REMOTE
202	6045	0.033	0.189	0.0063	REFS
195	1154	0.169	0.182	0.0308	SENSING
186	6340	0.029	0.174	0.0051	SYSTEM
185	1786	0.104	0.173	0.0179	INFORMATION
179	3073	0.058	0.167	0.0097	MEASUREMENTS

CODE:
Cij IS CO-OCCURRENCE FREQUENCY, OR NUMBER OF TIMES CLUSTER MEMBER APPEARS WITHIN +-50 WORDS OF CLUSTER THEME IN TOTAL TEXT; Ci IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER MEMBER; Cj IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER THEME; Ii, THE CLUSTER MEMBER INCLUSION INDEX, IS RATIO OF Cij TO Ci; Ij, THE CLUSTER THEME INCLUSION INDEX, IS RATIO OF Cij TO Cj, AND Eij, THE EQUIVALENCE INDEX, IS PRODUCT OF INCLUSION INDEX BASED ON CLUSTER MEMBER Ii (Cij/Ci) AND INCLUSION INDEX BASED ON CLUSTER THEME Ij (Cij/Cj).

In the following figures, the underlined topic is the cluster theme. The cluster members are segregated by their values of Inclusion Indices. Ij is the ratio of Cij to Cj, and is the Inclusion Index based on the theme word. Ii is the ratio of Cij to Ci, and is the Inclusion Index based on the cluster member. The dividing points between high and low Ij and Ii are the middle of the "knee" of the distribution functions of numbers of cluster members vs. values of Ij and Ii. All cluster members with Ij greater than or equal to approximately 0.1 were defined as having high Ij. All cluster members with Ii greater than or equal to 0.5 were defined as having high Ii. A high value of Ij means that, whenever the theme phrase appears in the text, there is a high probability that the cluster member will appear within +-50 words of the theme phrase. A high value of Ii means that, whenever the cluster member appears in the text, there is a high probability that the theme phrase will appear within +-50 words of the cluster member.

Thus, phrases located under the heading HIGH Ij HIGH Ii are coupled very strongly to the theme phrase. Whenever the theme phrase appears, there is a high probability that the cluster member will be physically close. Whenever the cluster member appears, there is a high probability that the theme phrase will be physically close. Whenever either phrase appears in the text, the other will be physically close.

Consider phrases located under the heading HIGH Ij LOW Ii. Whenever the cluster member appears in the text, there is a low probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the text, there is a high probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially larger than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning. Single word phrases have absolute frequencies of an order of magnitude higher than double word phrases. Thus, the phrases under the heading HIGH Ij LOW Ii are typically high frequency single word phrases. They are related to the theme phrase but much broader in meaning than the theme phrase. A small fraction of the time that these broad single word phrases appear, the more narrowly defined double word phrase theme will appear physically close. However, whenever the narrowly defined double word phrase theme appears, the broader related single word phrase cluster member will appear. The phrases under this heading can also be viewed as a higher level taxonomy of technical disciplines related to the theme.

Consider phrases located under the heading LOW Ij HIGH Ii. Whenever the cluster member appears in the text, there is a high probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the text, there is a low probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially smaller than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning. Thus, the phrases under the heading LOW Ij HIGH Ii tend to be low frequency double and triple word phrases, related to the theme phrase but very narrowly defined.

A large fraction of the time that these very narrow double and triple word phrases appear, the relatively broader double word phrase theme will appear physically close. However, a small fraction of the time that the relatively broad double word phrase theme appears, the more narrow double and triple word phrase cluster member will appear. This grouping has the potential for identifying "needle-in-a-haystack" type thrusts which occur infrequently but strongly support the theme when they do occur. One of many advantages of full text over key or index phrases is this illustrated ability to retain low frequency but highly important phrases, since the key phrase approach ignores the low frequency phrases.

REMOTE SENSING

The first grouping analyzed is the BLOCK database; low Ii high Ij. The words describe the more generic associations with REMOTE SENSING.

The major journals and conferences whose NES articles are associated with REMOTE SENSING include PROCEEDINGS OF SPIE, IGARSS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, REMOTE SENSING OF ENVIRONMENT, and INTERNATIONAL JOURNAL OF REMOTE SENSING [LATTER TWO ARE BOTH EC & SCI SOURCES-FIRST THREE ARE EC SOURCES].

The major generic types of sensors associated with REMOTE SENSING include RADAR SYSTEMS, OPTICAL DEVICES, RADIO, MICROWAVE, SYNTHETIC APERTURE RADAR, RADIOMETERS, INFRARED, ELECTROMAGNETIC, and MULTISPECTURAL.

The major applications associated with REMOTE SENSING include COMMUNICATIONS, MONITORING, METEOROLOGY, and MAPPING.

The major regions of interest associated with REMOTE SENSING include SPACE, ATMOSPHERIC, SURFACE, WATER ENVIRONMENT, SEA, LAND, MARINE, FOREST, and COASTAL.

The major variables of interest associated with REMOTE SENSING include TEMPERATURE,RADIATION, and VEGETATION, and the major generic products associated with REMOTE SENSING are IMAGE and INFORMATION.

Analysis tools associated with REMOTE SENSING include DATA PROCESSING, MATHEMATICAL MODELS, ALGORITHMS, COMPUTER SOFTWARE, IMAGEPROCESSING, and IMAGE ANALYSIS.

The second grouping analyzed is the BLOCK database; high Ii low Ij. The words describe the more specific associations with REMOTE SENSING.

Authors closely associated with REMOTE SENSING include CRACKNELL-AP, VARTSOS-CA, KONDRATEV-KY, GUSHIN-GA, ZAKHAROV-MY, LUPYAN-EA, STAELIN-D, BELCHANSKII-G1, CHUNG-LI, ERIC-R -SHUCHMAN -R, WACKERMAN-C, ALPERS-W, EYMARD-L, FRIEDL-MA, GOWARD- SN, GOWER-JFR, YAN-XH, SIMPON-JJ, LAMBIN-EF.

Journals closely associated with REMOTE SENSING in this database include PHOTOGRAMMATIC ENGINEERING, JOURNAL OF PHOTOGRAMMTRY, IGARRSS, and IEEE TRANSACTIONS (ON GEOSCIENCE AND REMOTE SENSING).

Institutions closely associated with REMOTE SENSING include UNIV-DUNDEE, INST. MARINE HYDROPHYS SEVASTAPOL UKRAINE, UNIV-DELAWARE, BOSTON UNIV, UNIV OF HAMBURG, UNIV-CALIF SANTA-BARBARA, UNIV-VALENCIA, UNIV-N- CAROLINA, UNIV-COLL-SWANSEA, IWATE UNIV, BAY-ST-LOUIS MS, SAGA UNIV, RUSSIAN-ACAD-SCI, UNIV MUNICH, UNIV OF MISSOURI, UNIV OF BERNE, and UNIV OF ZURICH.

Regions of interest closely associated with REMOTE SENSING include SENEGALESE SAHEL, ROSETTA PROMONTORY, RAIN FOREST, POLAR ICE, COASTAL ZONES, and GREENLAND ICE SHEET. Measuring instruments closely associated with REMOTE SENSING include MICROWAVE INSTRUMENTATION, MULTISPECTRAL SCANNERS RADIOMETERS, ENHANCED THEMATIC MAPPER, X-RAY, SAR, SCANNING MULTICHANNEL MICROWAVE, SATELLITE MICROWAVE RADIOMETER, SAR INTERFEROMETRY, AIRBORNE SPECTROGRAPHIC IMAGER, and RADIOMETERS INFRARED IMAGING.

Output quantities of interest associated with REMOTE SENSING include CANOPY REFLECTANCE MODEL, LEAF AREA INDEX, CROP YIELD, FORESTRY MAPPING, SEA SURFACE TEMPERATURES, SURFACE EMISSIVITY, VEGETATION DYNAMICS, ALPINE SNOW COVER, BIDIRECTIONAL REFLECTANCE DISTRIBUTION FUNCTION, CLIMATOLOGY DATA ACQUISITION, VEGETATED LAND SURFACES, WATER SURFACE TEMPERATURE, TERRAIN DATA BASES, LAND DEGRADATION ASSESSMENT, ACREAGE ESTIMATION, SOIL SALINITY, SNOWMELT RUNOFF, SNOW DEPTH, SEA-SURFACE VELOCITIES, SEA-ICE MAPPING, LAND USE MAPPING, RED TIDE, AGRICULTURAL STATISTICS, ANTECEDENT PRECIPITATION INDEX, CROP INDEX, CROP YIELD ESTIMATION, SURFACE CURRENT, OIL SPILLS, and PHOTOMAPPING SURVEYS.

The third grouping analyzed is the ABSTRACT database; low Ii high Ij.

Applications closely associated with REMOTE SENSING include MONITORING, MAPPING, DETECTION, and WEATHER.

Regions of interest associated with REMOTE SENSING include SURFACE, ATMOSPHERIC, SPACE, EARTH, SEA, LAND, OCEAN, EUROPEAN, INDIAN, COASTAL, SEA SURFACE, TROPICAL, and LAKE, while related features include VEGETATION, WATER, FOREST, SOIL, ICE, OIL, SNOW, CROP, WIND, CANOPY, and AGRICULTURAL.

Sensors closely associated with REMOTE SENSING include RADAR, SAR, SPECTRAL, MICROWAVE, OPTICAL, INFRARED, RADIOMETER, and MULTISPECTRAL, and the associated variables measured include IMAGE, TEMPERATURE, CURRENT, SOIL MOISTURE, AEROSOL, SURFACE TEMPERATURE, and RAIN.

The fourth grouping analyzed is the ABSTRACT database; high Ii, low Ij.

Applications closely associated with REMOTE SENSING include DETECTION OF OIL SLICKS, OIL SLICK TRACKING, MONITORING FREEZE-THAW CYCLES, VEGETATION MAPPING, and TRACKING USING DRIFTING BUOYS AND AIRCRAFT.

Regions of interest closely associated with REMOTE SENSING include MARINE ATMOSPHERIC SURFACE LAYER, COASTAL ENVIRONMENTS, TROLLFJORD-KOMAGELV FAULT ZONE, VARANGER PENINSULA, AURORAL ZONES, TERRESTRIAL ECOSYSTEMS, and associated features of interest include SURFACE MINING, WHEAT ACREAGE, DARK DENSE VEGETATION, SNOW HYDROLOGY, COAL MINING, CORAL REEF, SEA SURFACE MANIFESTATIONS, OBLIQUE SHEARS, UNSTRESSED CANOPY, and BLACK SPRUCE PICEA MARIANA.

Sensors closely associated with REMOTE SENSING include SCANNING APERTURE RADAR, RADAR SAR IMAGERY, SYNTHETIC APERTURE SONAR, COMPACT AIRBORNE SPECTROGRAPHIC IMAGER, SATELLITE RADAR ALTIMETRY, ACTIVE MICROWAVE INSTRUMENT, OPTICAL AND MICROWAVE, and THERMAL INFRARED.

Their associated measured variables include SHORE DRIFT DIRECTION, SPECTRAL REFLECTANCE DATA, EARTH RESOURCES DATA, MARINE POLLUTION, SOIL HEAT FLUX, PARTICLE SIZE DISTRIBUTION, REFLECTANCE DISTRIBUTION FUNCTION, and SUSPENDED SOLIDS CONCENTRATION.

The output quantities derived from these measurements include ATMOSPHERIC COMPOSITION, WHEAT YIELD, ENVIRONMENTAL IMPACTS, NET CO2, FOREST PARAMETERS, LAND-SURFACE ENERGY BALANCE, AGRICULTURAL STATISTICS, COHERENT CONVECTIVE STRUCTURES, OCEAN MICROWAVE PROPERTIES, VEGETATION COVER MAPS, CROP CLASSIFICATION, GLOBAL CARBON CYCLE, SEWAGE OUTFALL, ENVIRONMENTAL DEGRADATION, SEA ICE MAPS.

Data analysis methods include DIGITAL IMAGE PROCESSING and AIR PHOTO INTERPRETATION.

TAXONOMIES

The different types of Database Tomography outputs allow different types of taxonomies, or classifications into component categories, to be generated. Such categorizations, analogous to the independent axes of a mathematical coordinate system, allow the underlying structure of a field to be portrayed more clearly, leading to more focused analytical and management analyses. Three separate taxonomies will be discussed here, but due to space limitations, only one will be presented. The first taxonomy derives from the phrase frequencies. The authors examined the phrase frequency outputs, then arbitrarily grouped the high frequency phrases into different, relatively independent, categories for which all remaining terms would be accounted. One taxonomy was developed for the SCI phrase frequencies, and another taxonomy for the EC phrase frequencies. The second taxonomy derives from the phrase frequency and proximity analysis. From the phrase frequency analyses, fifty or sixty high frequency technical phrases were identified as pervasive themes. The next step was to group these high frequency phrases into categories of related themes. A proximity analysis was done for each of these high frequency phrases. A word frequency dictionary, or cluster, was generated for each phrase. This cluster contained those phrases which were in close physical proximity to the pervasive theme throughout the text. The degree of overlap among clusters was computed. Clusters which shared more than a threshold number of common phrases were viewed as overlapping. These overlapping clusters were viewed as links in a chain, with the different chains being relatively independent. Each chain was then defined as a category of the larger taxonomy. The third taxonomy derives also from the phrase proximity analyses. A cluster was formed for SATELLITE(S), and the phrases closely related to SATELLITE in the joint SCI/ EC database were identified and ranked. The phrases were separated into relatively independent categories to form a taxonomy.

TABLE 14 - TAXONOMIES

<u>TAXONOMY 1A - SCI - PHRASE FREQUENCY BASED</u>

*SPACE PLATFORM (E.G., SATELLITE, SPACECRAFT)
*SATELLITE FUNCTION (E.G., MAPPING, NAVIGATION)
*SATELLITE TYPE (E.G., GEOSAT, LANDSAT)
*MEASURING INSTRUMENT (E.G., RADIOMETER, MICROWAVE IMAGER)
*REGION EXAMINED (E.G., SEA, BOUNDARY LAYER)
*LOCATION EXAMINED (E.G., NORTH ATLANTIC, SOUTHERN HEMISPHERE)
*VARIABLE MEASURED (E.G., TEMPERATURE, SOIL MOISTURE)
*VARIABLE DERIVED (E.G., RADIATION BUDGET, GENERAL CIRCULATION)

*ANALYTICAL TOOL (E.G., DATA PROCESSING, MATHEMATICAL MODELS)
*PRODUCTS (E.G., TIME SERIES, SEA ICE MAPS)
*SPACE ENVIRONMENT (E.G., SOLAR WIND, MAGNETIC FIELD)

TAXONOMY 1B - EC - PHRASE FREQUENCY BASED

SAME AS 1A, BUT ADD:

*SATELLITE CONFIGURATION (GEOSTATIONARY SATELLITES, TETHERED SATELLITE SYSTEM)
*SATELLITE STATE (ATTITUDE DETERMINATION, HIGH ELEVATION ANGLE)
*SATELLITE SUBSYSTEMS (SOLAR CELLS, ATTITUDE CONTROL SYSTEM)

CONCLUSIONS AND APPLICATIONS

This report has provided a map of the NES field, although only a small fraction of the raw data has been presented. A Competitive Intelligence (CI) professional who has interest in this field has many options for proceeding further from the map, depending on this person's specific interests. For example, if the analyst wanted to understand the intellectual foundations of NES, then a reading of the most highly cited papers would be an excellent starting point. If the analyst wanted to overview the current literature, then two approaches are available. The comprehensive literature survey used as the database for this analysis is one avenue. Another is to peruse the journals which contain the highest frequency of recent publications. This latter approach is worthwhile since computerized search approaches don't always identify the full scope of related articles to the topic of interest, and journals which focus on such a topical area could yield a cornucopia of useful information through browsing. If the analyst wants to contact experts in a particular thrust area or technique, then contact could be made with the specific individuals or the institutions identified with given techniques in the theme relationships section. The key conclusion is that, starting from the raw data, the analyst can generate any cross-cutting relationships desired to proceed further in specific directions of personal interest.

REFERENCES

1. Kostoff, R. N., et al, "System and Method for Database Tomography", U.S. Patent Number 5440481, August 8, 1995.
2. Kostoff, R. N., "Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis," Proceedings: Portland International Conference on Management of Engineering and Technology, October 27-31, 1991.
3. Kostoff, R. N., "Database Tomography for Technical Intelligence," Competitive Intelligence Review, 4:1, Spring 1993.
4. Kostoff, R.N., "Database Tomography: Origins and Applications," Competitive Intelligence Review, Special Issue on Technology, 5:1, Spring 1994.
5. Kostoff, R. N., "Database Tomography for Technical Intelligence: Analysis of the Research Impact Assessment Literature, Competitive Intelligence Review, Fall, 1997.
6. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", Journal of Information Science, 23:4, 1997.
7. Kostoff, R. N., "The Handbook of Research Impact Assessment", Sixth Edition, Summer 1996, DTIC Report Number ADA296021, Fort Belvoir, VA 22060.